SparQLe: Speech Queries to Text Translation Through LLMs

Amirbek Djanibekov, Hanan Aldarmaki
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, UAE
{amirbek.djanibekov;hanan.aldarmaki}@mbzuai.ac.ae

Abstract

With the growing influence of Large Language Models (LLMs), there is increasing interest in integrating speech representations with them to enable more seamless multi-modal processing and speech understanding. This study introduces a novel approach that combines self-supervised speech representations with instruction-tuned LLMs for speech-to-text translation. The proposed approach leverages a modality adapter to align extracted speech features with instruction-tuned LLMs using English speech data. Our experiments demonstrate that this method effectively preserves the semantic content of the input speech and serves as an effective bridge between self-supervised speech models and instruction-tuned LLMs, offering a promising approach for various speech understanding applications.

1 Introduction

Progress in speech processing has been accelerated by the introduction of self-supervised learning (SSL) methods that utilize large amounts of unlabeled speech data, which established new benchmarks in the field (Xu et al., 2021; Hsu et al., 2021; Zeghidour et al., 2021). Continuous representations and/or discrete units derived from self-supervised models have been used to extract relevant latent features from speech data and improve performance in downstream tasks, including speech recognition (Baevski et al., 2020), speech synthesis (Ren et al.; Wang et al., 2023b), speech translation (Inaguma et al., 2020) and general speech understanding (Wang et al., 2020). Progress in text processing has also been accelerated by the emergence of pre-trained Large Language Models (LLMs), which enabled new applications such as few-shot/zero-shot language processing (Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023; Bai et al., 2023) and multi-modal processing (Tsimpoukelli et al., 2021; Radford et al., 2021). Recent efforts in speech understanding explored the possibility of incorporating speech representations directly into LLMs (Zhang et al., 2023; Wang et al., 2023c; Das et al., 2024; Fang et al., 2024). Multi-modal speech-language models signify a shift in both speech and natural language processing. By incorporating speech data, LLMs can enhance their contextual grasp, providing a deeper and more thorough representation of spoken language. In addition, the existing multilingual functionalities of LLMs can be leveraged to enhance speech processing applications, such as speech translation, without additional dedicated training.

In this work, we describe an efficient method to query instruction-tuned LLMs using speech input. The model aligns speech features extracted through self-supervised learning (SSL) with LLMs using only a modality adapter trained with English data and a small portion of translated text. We demonstrate the generalization of translation performance across both seen and unseen target languages. We call our approach SparQLe [1]. SparQLe is inspired by Querying Transformer modules used in vision language models to bootstrap vision-language representations from frozen image encoders (Li et al., 2023). We demonstrate through speech translation that SparQLe enables the integration of existing pre-trained speech encoders and LLMs without the need for updating the parameters of either speech encoder or LLM. In contrast to previously explored speech-LLM integration approaches, our method is the first to utilize frozen SSL speech representations, without relying on large pre-trained ASR models like Whisper (Radford et al., 2023). We experimentally demonstrate the effectiveness of this relatively simple approach and release both the pre-trained and fine-tuned models [2].

[1] **Sp**eech **R**outing to **Q**uery Large **L**anguage models.
[2] https://github.com/djanibekov/rebooting-llm

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 76–83
July 31 - August 1, 2025

Method	Speech Encoder (Param.)	Language Model (Param.)	Adapter (Param.)	Tasks
(Chen et al., 2024)	NeMo (0.6B - 1.1B)	MegatronLLM (40B-1T)	LoRA (14M-94M) / Conformer (115M)	multitask
(Wang et al., 2023a)	Whisper (74M-1.5B)	LLama (6.7B-65.2B)	Conv.Layers (4M)	alignment
(Wang et al., 2023c)	USM (2B)	mT0-MT XXL (13B)	Adapter (156M)	multitask
(Wang et al., 2023d)	CTC encoder (220M)	T5 XXL (11B) - RAG	Speech2Text, Speech2Entity Retriever	multitask
(Yu et al., 2024)	Whisper (1.5B)	VicunaLLM (13B)	FC[3] (24M) / MHSA[4] (133M) / Seg-Q-Former (24M)	ASR
(Tang et al., 2024)	Whisper (1.5B) + BEATS (90M)	VicunaLLM (13B)	LoRA + Seg-Q-Former (33M)	ASR/multitask
(Das et al., 2024)	WavLM (316.62M)	Flan-T5-XL (2.85B)	CNN + LoRA(14M-94M)	multitask
(Chu et al., 2024)	Whisperlarge (1.5B)	Qwen (7B)	—	multitask
SparQLe	HuBERT (316M)	LLama3 (8B)	Q-Former (187M)	AST/multitask

Table 1: Comparison of related works and proposed model. LoRA's rank in (Chen et al., 2024) is assumed to be 8. For other rank values, multiply number of parameters by 2 for 16 and 4 for 32 ranks, respectively.

2 Related Works

The availability of instruction-tuned LLMs (Touvron et al., 2023; AI@Meta, 2024; Jiang et al., 2023) opened a new research direction for speech processing by connecting speech directly to these multi-task models. Chen et al. (2024) proposed multitask speech-language modeling with unified LLM framework that shows in-context learning ability. Yu et al. (2024) utilized three approaches for adapting speech to text modality: Fully Connected Linear Layers following (Houlsby et al., 2019) adapter method, multi-head cross attention mechanism described in (Vaswani et al., 2017), and query transformer (Li et al., 2023). For processing speech input, they utilized two models: Whisper Large-v2 (Radford et al., 2023) and BEATS (Chen et al., 2023). The SpeechVerse (Das et al., 2024) framework used WavLM-based (Chen et al., 2022) speech encoder interfaced with a Flan-T5-XL (Chung et al., 2024) language model. In a another study, Ma et al. (2024) demonstrated the sufficiency of a single linear layer for speech-LLM integration in ASR, albeit with limited exploration beyond this task. Speech as language modeling was also studied in SpeechGPT (Zhang et al., 2023) which integrates both speech and text modalities. The model incorporates a speech tokenizer that converts raw audio waveforms into discrete speech tokens, enabling efficient processing within the transformer architecture. Through multi-task fine-tuning on downstream tasks such as ASR, translation, and generation, the model demonstrates remarkable versatility. Qwen2-Audio (Chu et al., 2024), designed as a general-purpose audio understanding model, exhibiting broad applicability across various audio-related tasks. The model employs self-supervised learning techniques, such as masked audio modeling and contrastive learning, to capture rich audio representations.

Table 1 summarizes the features of most relevant related works. We outline that, depending on the rank of the LoRA (Hu et al., 2021) adapter, the final number of trainable parameters can increase. LoRA rank is the number of linearly independent rows or columns in a parameter (weight) matrix; a lower rank means approximating a large weight matrix with fewer parameters to simplify and speed up training. The number of additional parameters can be roughly estimated as the initial hidden dimension multiplied by the rank and then multiplied by two to account for all added parameters. In Table 1, we outline the range of the possible numbers of added parameters. Note that our proposed model, SparQLe, is the only one that relies exclusively on SSL features (i.e. HuBERT) as input, and a simple adapter between the frozen speech encoder and LLM; previous approaches relied on complex encoders that have already been aligned with text through supervised training or adapt selected LLM with LoRA adapter.

3 Model

SparQLe is a parameter efficient model designed to extract information from speech representation and route them to query pre-trained open-sourced LLMs, without modifications to the underlying speech encoder or LLM. Motivated by the success of multi-modal representations in vision language modeling (Li et al., 2023), we propose the adoption of speech representations to LLMs for generative tasks, specifically Automatic Speech Translation (AST). We pre-trained our model using English data first and fine-tuned with mix of English and French.

We used HuBERT (Hsu et al., 2021) as the speech encoder. The output from its final hidden layer is fed into the query adapter. A query adapter incorporates query tokens, which are special tokens (placeholders) added to the input of a speech-language model. They do not correspond to specific

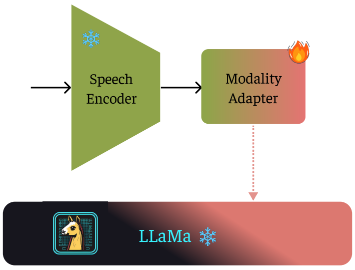

Figure 1: High-level overview of the SparQLe model

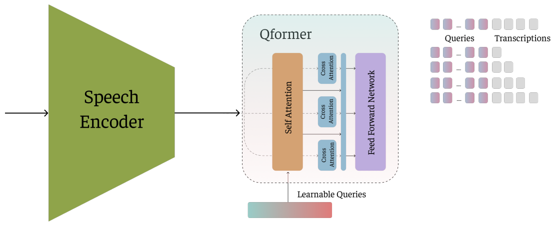

Figure 2: Modality adapter with auto-regressive supervised fine-tuning phase. The modality adapter is the Q-Former, which we discuss in the paper.

speech regions, but are meant to extract information from the whole speech sequence in a flexible way. The final features from the query tokens are passed to a large language model to generate natural language responses. Figure 1 shows the overall structure of the system, which consists of three main parts: a pre-trained speech encoder, a bridging mechanism (SparQLe) and a text generator. In our experiments, we employed LLama3 (AI@Meta, 2024) as the text language model.

3.1 Pre-Training

This stage is akin to ASR training, where we utilize transcribed speech for supervised training. However, we do not introduce additional parameters and instead use the same modality adapter as an auto-regressive language model: each output vector from the Q-Former (Li et al., 2023) is successively fed into a modality adapter to predict the next token. We only update the parameters of the adapter, and keep the underlying speech encoder frozen. The Q-Former is a vanilla transformer model but with learnable query tokens. These tokens are randomly initialized and designed to be learned during training to capture query information that is relevant to the task. The process is depicted in Figure 2. In addition to the text generation task, we use various modality alignment objectives to account for speech in the input and aligned text-like features in the output, similar to image-text alignment done in Li et al. (2023): **Speech-text contrastive learning** aligns speech and text representation such that mutual information is maximized. This is achieved through contrasting speech-text cosine similarity of positive against negative pairs. **Speech-text matching** loss aligns representations of speech and text via a binary classification task. **Speech text generation** loss trains the model to produce text based on the given audio.

3.2 Fine-tuning

After pre-training, we fine-tune the adapter on downstream tasks using an instruction-tuned LLM, specifically LLama3. We utilize the extracted query tokens as the input to the LLM and update the adapter parameter using cross entropy loss derived from the LLM's objective. For instruction tuning, a frozen Large Language Model was fed with a randomly selected pool of prompts, which were designed to define the translation task. Subsequently, the instruction-tuned model was employed in a chat-based format to collect predictions.

4 Experiments

4.1 Datasets

To train and evaluate the model on Automatic Speech Translation (AST) task, we used the MuST-C (Di Gangi et al., 2019) and LibriSpeech (Panayotov et al., 2015) datasets. Specifically, we selected the French and German languages from MuST-C for AST evaluation. We normalized the text across datasets by converting all letters to lowercase and eliminating punctuation marks. The MuST-C dataset includes action descriptions within the audio samples, such as "`<|speech|> (applause) <|speech|>`", which signify auditory sequences where spoken content is interspersed with audience applause. We opted to remove these actions from the translation text.

4.2 Pre-Training

4.2.1 Experimental settings

For feed-forward networks and self-attention of the modality adapter, we employ a 12-layer transformer-based Q-Former that is, by design, a UniLM (Dong et al., 2019); cross-attention is initiated randomly. For pre-training experiment, we trained the model using Adam optimizer, coupled with a cosine annealing learning rate scheduler during the pre-training. The learning rate was initiated

Figure 3: Sample of zero-shot instruction generation across multiple languages. To evaluate zero-shot capability, we simply changed the output language specified in the prompt. The produced text is lowercased and punctuation-free, following the text processing guidelines described in Section 4.1.

at 1×10^{-4} and gradually reduced to 1×10^{-5}, incorporating a warm-up phase at 1×10^{-6}. This means that learning rate started with warmup value and gradually reached from 10^{-6} to 10^{-4}. The maximum length for speech samples was capped at 480K frames, which is equivalent to 30s of audio. We used 100 learnable query tokens in Q-Former.

Our experiments were conducted using open-source library for language-vision intelligence, LAVIS[5]. The training processes were executed on one RTX4090 GPU with 24G memory being used over a period of two-three weeks with batch size equal to 8 due to memory constraints.

4.3 Fine-Tuning

4.3.1 Experimental Settings

We used 960 hours of audio from the LibriSpeech dataset, along with an additional 457 $\times$ 2 hours of audio samples from MuST-C that included both translation and transcription tasks: 70% of the speech samples for fine-tuning were used for recognition, while the remaining 30% involve English-to-French translation. We deliberately restricted our training data to one language in order to demonstrate the capacity of the model to generalize to other languages[6].

4.4 Prompts

We derived instruction prompts from SALMONN's (Tang et al., 2024) work. As demonstrated, each prompt includes a placeholder for speech `<Speech><SpeechQuery></Speech>`, into which we insert query-extracted embeddings as inputs to the LLM. Please note that the query embeddings are placed inside the placeholder denoted by `<SpeechQuery>`. Here is the prompts that we used for training:

[5] https://github.com/salesforce/LAVIS

[6] Instruction tuning sometimes results in overfitting to the training instructions, as observed in Tang et al. (2024).

	MuST-C_En-Fr BERTScore ↑	MuST-C_En-De BERTScore ↑
STRONGBASELINE	81.75%	77.44%
WEAKBASELINE	77.28%	74.86%
SparQLe	**85.56%**	**83.26%**

Table 2: Comparison of SparQLe against strong and weak baselines from the IWSLT isometric speech challenge (Anastasopoulos et al., 2022).

- `<Speech><SpeechQuery></Speech>` Can you translate the speech into "Language"?
- `<Speech><SpeechQuery></Speech>` Please translate the speech you heard into "Language".
- `<Speech><SpeechQuery></Speech>` Listen to the speech and translate it into "Language".
- `<Speech><SpeechQuery></Speech>` Give me the Language translation of this "Language".

`"Language"` can be any language a user wants to add to instruction.

4.4.1 Results & Analysis

We benchmarked translation against the IWSLT challenge baselines for speech-to-text translation using BERTScore (Zhang* et al., 2020) as reported in (Anastasopoulos et al., 2022). The results for English-German translation are zero-shot since the model is only fine-tuned with English-French speech translation data. Evaluating LLM answers for speech translation is a challenging task, primarily due to the presence of chat-specific artifacts in the output, such as prompt repetition, follow-up comments, and connecting phrases (e.g., `"here is the transcribed text:"`). To address this issue, we implemented a post-hoc approach in which we endeavored to eliminate instances of prompt recurrence (chat artifacts) in the final text. We consider two baseline systems from IWSLT2022 campaign (Anastasopoulos et al., 2022): **WEAKBASELINE** refers to a standard neural machine translation model trained under limited data conditions, without incorporating any isometric translation features. **STRONGBASELINE** is trained using unconstrained data setting and incorporates output length control following the approach of Lakew et al. (2021). This method involves adding a length token at the beginning of the input, generating N-best candidate translations, and then reranking them based on a weighted combination of the model's score and the length ratio.

The results in Table 2 highlight the generalization potential of the model in translation. Specifically, the BERTScore for the tst-COMMON split in the French language demonstrates that our system has surpassed both the WEAKBASELINE and STRONGBASELINE in terms of semantic similarity. Furthermore, evaluations on the tst-COMMON split for the German language show that the performance quality extends to languages not included in the training set. This success can be attributed to the inherent translation performance of the underlying LLM, demonstrating the model's adaptability to new instructions.[7].

5 Discussion

We introduced a framework for efficient routing of SSL speech features to query LLMs, and demonstrated its effectiveness in speech translation tasks. The results indicate that the proposed model and training paradigm result in generalized performance and avoid instruction over-fitting; the model was able to adhere to instructions for translating speech into multiple target languages (see Figure 3). SparQLe demonstrates ability to translate speech input into diverse languages not encountered during our fine-tuning stage, such as German, Russian, Arabic, etc. Finally, with the appropriate prompts, the instruction-tuned model is capable of performing multiple tasks in a single run, (see Figure 4 in Section 5.1). The performance in speech translation shows promising results, where the proposed approach outperformed both weak and strong baselines from Anastasopoulos et al. (2022) in both French and German.

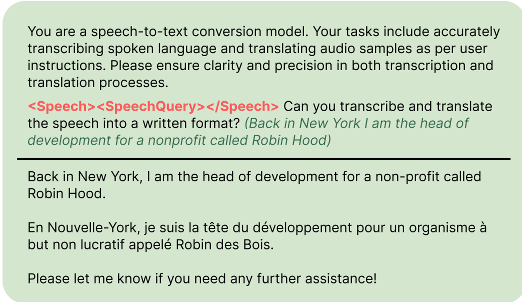

Figure 4: Example from the SparQLe for multi-tasking in one prompt.

5.1 Multi task Discussion

As mentioned before with the appropriate prompts, the instruction-tuned model is capable of performing multiple tasks in a single run, See Figure 4. While we have not conducted an exhaustive analysis of this aspect in the current study, this example illustrates potential applications for efficiency and versatility in spoken language applications.

6 Conclusion & Future Work

In this short paper, we demonstrate the performance of the proposed SparQLe model, an aligned speech-to-text model based on SSL features, for speech translation applications. What we have demonstrated in this study is only a subset of potential applications of this method. The SparQLe model can potentially handle both text and speech modalities, and can be applied for any speech-to-text applications. As demonstrated, our model outperforms existing speech translation baselines from IWSLT 2022 challenge, which demonstrates the potential of transferring the inherent capacities of LLMs into speech tasks using a parameter-efficient approach. Future work can explore the generalization of the model to other languages and speech understanding tasks and analyze the characteristics of the resulting queries.

[7]During the inference phase, we executed four different prompts which are listed in Section 4.4. We selected the prompt that yielded the best results on held-out set.

Limitations

Our model was initially pre-trained to align specifically with English speech samples, disregarding other rich languages that present unique challenges. While we believe SparQLe has the potential to handle various tasks beyond its original training scope, we have not yet carried out a formal assessment to verify this capability. Although our model is adaptable to multiple LLMs, we onlye explored one model. Similarly, we did not explore other speech encoders apart from HuBERT. For translation evaluation we used BERTScore, which measures semantic similarity for generation tasks, but all automatic translation metrics have limitations. For example, sentences `"never had any act seemed so impossible"` and `"always had any act seemed so impossible"` convey different information but are similar in words. BERTScore outputs that these two sentences have a high similarity score, which is, in fact, not true (99.7% in F1 score). We did not test our model on tasks other than translation and transcription. As a result, the model's performance on other modalities or tasks, such as speech question answering, remains unverified.

References

AI@Meta. 2024. Llama 3 model card.

Antonios Anastasopoulos, Loc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, and 1 others. 2022. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525. IEEE.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, and 1 others. 2024. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. Espnet-st: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Surafel M Lakew, Marcello Federico, Yue Wang, Cuong Hoang, Yogesh Virkar, Roberto Barra-Chicote, and Robert Enyedi. 2021. Machine translation verbosity control for automatic dubbing. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7538–7542. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. 2024. An embarrassingly simple approach for llm with strong asr capacity. *Preprint*, arXiv:2402.08846.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2023a. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023b. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul K Rubenstein, and 1 others. 2023c. Slm: Bridge the thin gap between speech and text foundation models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Mingqiu Wang, Izhak Shafran, Hagen Soltau, Wei Han, Yuan Cao, Dian Yu, and Laurent El Shafey. 2023d. Speech-to-text adapter and speech-to-entity retriever augmented llms for speech understanding. *arXiv preprint arXiv:2306.07944*.

Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, and 1 others. 2020. Transformer-based acoustic modeling for hybrid speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6874–6878. IEEE.

Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034. IEEE.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Connecting speech encoder and large language model for asr. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12637–12641. IEEE.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.